A simplified representation of anisotropic charge distributions within proteins

Travis Hoppe^{a)}

National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland 20892, USA

(Received 24 January 2013; accepted 15 April 2013; published online 3 May 2013)

Effective coarse-grained representations of protein-protein interaction potentials are vital in the modeling of large scale systems. We develop a method to fit an arbitrary number of effective charges to approximate the electrostatic potential of a protein at a given pH in an ionic solution. We find that the effective charges can reproduce an input potential calculated from a high resolution Poisson-Boltzmann calculation. Since the effective charges used in this model are not constrained to the locations of the original charged groups, the extra degrees of freedom allows us to reproduce the field anisotropy with fewer charges. The fitting procedure uses a number of approximations in the charge magnitudes, initial conditions, and multipoles to speed convergence. The most significant gains are found by fitting the multipole moments of the effective charge potential to the moments of the original field. We show that the Yukawa potential is not only sufficient as a pairwise summation in reproducing the potential, but comes naturally from the linearized expansion of the Poisson-Boltzmann equation. We compute interaction energies and find excellent agreement to the original potential. From the effective charge model we compute the electrostatic contribution to the second virial coefficient. [<http://dx.doi.org/10.1063/1.4803099>]

I. INTRODUCTION

A simple yet effective representation of a protein is important when modeling large systems of protein-protein interactions. Hard-spheres are the simplest such models. While they only capture the excluded volume effect, a wide range of isotropic interactions can be modeled by fitting an effective radius.¹ More complex isotropic potentials, such as the square-well or Lennard-Jones, are necessary² and sufficient³ to capture liquid-liquid phase transitions. While these simple representations are amenable to theoretical and computational calculations,^{4,5} they fail to fully account for one of the major components, electrostatic interactions. At this level of coarse-graining, electrostatic interactions are usually modeled as point monopoles or dipoles.^{6,7} As such, the capacity of these systems to account for a large degree of anisotropy is limited. While this is sufficient when the monopole moment is the dominant interaction, this failure is especially apparent when the protein has higher order multipole moments or when the solution is near an isoelectric point.

In contrast, all-atom *ab initio* calculations attempt to capture more detail by including explicit solvent representations or even models of the electron density. While that approach is more accurate, the computational burden becomes intractable when the number of components becomes large.

The goal of this paper is to merge a combination of the two methodologies. We attempt to mimic the more accurate high resolution electrostatic potentials with a simple coarse-grained representation of effective charges. Once found, this representation can be used in studies of protein-protein in-

teractions with a greater number density of molecules than currently possible in all-atom models.

There are two dominant methods to capture the salient features of protein electrostatic interactions, multipole methods, and effective/macro charges. Multipole methods use an expansion of the underlying potential as a basis set of the underlying differential equations. For Coulomb potentials of the $1/r$ type, the expansion is a series of spherical harmonics. The expansion and implementation of $1/r$ potentials for theoretical and numerical simulations is well developed.^{8–10} When screening effects are present, such as those that arise from the shielding of protonated charges in an ionic solvent, a Yukawa potential of the e^{-kr}/r type is a more realistic representation of the distance dependence. For Yukawa potentials, the expansion is a combination of spherical harmonics and modified spherical Bessel functions. Regardless of the expansion, the terms are truncated at some low finite order.¹¹ Intra-protein interactions are reduced to the interaction of multipole moments. The computation of this interaction, however, can be quite complicated. The so-called Fast Multipole Method (FMM) can be used to compute the energy of two separate interacting multipole expansions.^{12,13} The implementation itself can be complicated as it requires the repeated translation of far-field multipole expansions to local ones.

Effective charge methods seek to replace the electrostatic potential (ESP) by a set of discrete macrocharges whose individual contributions mimic the original potential. The input ESP can be calculated for a fixed conformation to high-precision when the solvent is approximated via the Poisson-Boltzmann (PB) equation. If inter-protein interactions dominate, effective charges can serve as a good surrogate to the real potential when studying structural properties of protein

^{a)}hoppeta@mail.nih.gov

solutions. However, for intra-protein interactions both charge location and magnitude can fluctuate wildly over small conformational changes found in flexible molecules.^{14,15} As such, the effective charge model is a poor representation for intra-protein interactions and if used in a numerical simulation it must be calculated by other means.

Prior implementations of effective charge models placed the locations of the charges at the locations of original charged groups or on the surface of a sphere.^{14–17} Since the potential is linear in charge magnitude, a unique solution can be found by fixing the charge positions. While the resulting effective charge potential gave a good fit, the reproduction required at least as many charges as the number of charged residues on the original protein. For example, in human serum albumin, a protein studied in this paper, there are approximately a hundred charged residues. In the computation of a pairwise potential, the number of evaluations scales quadratically with the number of charges. From a computational standpoint, if the same accuracy can be achieved with a smaller number of charges, larger systems can be modeled. In this paper we investigate fitting the potential to a subset of the original charge number. To compensate for the degrees of freedom removed we let the effective charges take arbitrary positions inside the protein with only a small number of constraints.

An approach similar in spirit, the Discrete Surface Charge Optimization (DiSCO) algorithm, was pioneered by the Schlick group.^{18,19} In their model, the potential was coarse-grained to a linearized sum of multiple Debye-Hückel spheres. While the objective of finding a computationally tractable potential is the same, our approach differs in two aspects. First, our computational method relies on a spherical harmonic and modified Bessel function decomposition to optimize the positions of the effective charges, a technique that could be applied to their model as well. Second, the domain of our effective charge distribution includes the entire accessible volume and is not limited to a surface. The choice of a surface domain may be more appropriate for the geometrically irregular objects studied in their works, e.g., DNA. In this paper we concentrate on modeling spherical-like domains found in globular proteins. While the choice of a volume domain substantially increases the parameter search space, we found that it allows for a more compact representation resulting in fewer effective charges.

Why should we expect that a subset of charges reproduce the ESP at all? In the simplest case, where the electrostatic potential is far from the isoelectric point, the potential is simply that of a monopole. In this case a single equivalent charge will account for the majority of the field. As one approaches the isoelectric point, the monopole moment vanishes and the anisotropy increases. Nonetheless, we will show that given a good basis set, a small number of higher-order multipoles can account for almost all of the field.

This paper attempts to combine the strengths of both the effective charge model and a multipole representation. Stated simply, we want to solve the inverse problem: given the value of the ESP as a function of spatial coordinates in a volume surrounding a protein molecule, what effective charges placed within the protein will best reproduce the spatial dependence of the ESP? The full potential, representing the interaction of

charges in an inhomogeneous dielectric medium, will be replaced with a set of discrete point charges in a homogeneous medium whose interaction is screened. The fit of these effective charges is not limited to the known charge locations or artificially constrained to a particular geometry. Instead we fit the charges in a free configuration at the cost of making the process nonlinear. We offset this cost by fitting the multipole moments of the two distributions rather than fitting the fields directly.

While this paper relies on the approximations of the Poisson-Boltzmann equation for the description of the ionic solution, we note that our method is intended to solve the inverse problem for screened interactions in general. Thus the methodology should be applicable for any potential that can be well approximated by a discrete sum of screened Coulombic interactions. Effective charge models have been found in the context of charge renormalization theories regarding spherical, elliptical or rod-like colloidal suspensions.^{20–22} These models, which typically match charges in a Wigner-Seitz sphere, consider interactions between charged colloidal particles that are sufficiently strong to cause crystallization. Our method is similar in spirit, in that we only consider the Coulombic interactions as the dominant potential. This treatment precludes an accurate description of short-ranged dispersion forces and ion-ion interactions (other than the effects of finite ion size). Nevertheless, the results will still be a valid representation of the inter-protein electrostatic potential at the mid to far field range.

This paper is organized as follows. In Sec. II we review the theory of the multipole moment expansions of the Coulomb and Yukawa potentials. In Sec. III we describe our fitting procedure. Our results are given in Sec. IV, where we discuss how well the effective charges compare to those generated by the initial potential. We use these effective charges to compute interaction energies and virial coefficients. Finally, in Sec. V we discuss how the model can be implemented in a wider context and to greater use.

II. THEORY

In this section we briefly describe the multipole expansions of both the Coulomb and Yukawa potentials. For further details see Refs. 23 and 24. We also review the construction of the Poisson-Boltzmann equation. It is important to note that the Yukawa potential is recovered from the Poisson-Boltzmann equations after linearization.

A. Multipole expansions

For a system of N discrete charges located at $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ with charge magnitudes $\mathbf{q} = [q_1, q_2, \dots, q_N]$ in a uniform medium, the Coulomb potential up to a constant at a location $\mathbf{r}$ is

$$\Phi(\mathbf{r}) = \sum_i^N q_i \frac{1}{\|\mathbf{r} - \mathbf{r}_i\|}. \quad (1)$$

The governing interaction $1/r$ is the Green's function for the partial differential equation $\nabla^2 \Phi = \rho(\mathbf{r})$, where $\rho(\mathbf{r})$ is an

arbitrary charge distribution. This potential can be expanded as a series of spherical harmonics

$$\frac{1}{\|\mathbf{r} - \mathbf{r}_i\|} = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \frac{4\pi}{2\ell+1} \frac{r_{<}^{\ell}}{r_{>}^{\ell+1}} Y_{\ell m}^*(\hat{\mathbf{r}}_i) Y_{\ell m}(\hat{\mathbf{r}}). \quad (2)$$

Here $Y_{\ell m}$ denotes a spherical harmonic of degree ℓ and order m , $Y_{\ell m}^*$ represents complex conjugation, a hat denotes a unit vector $\hat{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|$, and $r_{<} = \min(\|\mathbf{r}\|, \|\mathbf{r}_i\|)$, $r_{>} = \max(\|\mathbf{r}\|, \|\mathbf{r}_i\|)$.

In ionic solutions, the charges are partially screened by the solvent. A first order approximation to this screening can be expressed through the differential relation $(\nabla^2 - \kappa^2)\Phi = \rho(\mathbf{r})$. The inverse Debye length κ , acts a screening parameter (see Sec. II B for more details). Since the Green's function for this equation is $e^{-\kappa r}/r$, a charge distribution that can be expressed as a sum of discrete charges leads to the Yukawa potential

$$\Phi(\mathbf{r}) = \sum_{i,j \neq i}^N q_i \frac{e^{-\kappa \|\mathbf{r} - \mathbf{r}_i\|}}{\|\mathbf{r} - \mathbf{r}_i\|}. \quad (3)$$

The Yukawa potential can be expanded as

$$\frac{e^{-\kappa \|\mathbf{r} - \mathbf{r}_i\|}}{\|\mathbf{r} - \mathbf{r}_i\|} = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} i_{\ell}(\kappa r_{<}) k_{\ell}(\kappa r_{>}) Y_{\ell m}^*(\hat{\mathbf{r}}_i) Y_{\ell m}(\hat{\mathbf{r}}), \quad (4)$$

where i and k are the modified spherical Bessel functions. See Appendix A for a more complete description of these terms. If the charges are located inside a sphere of radius a , centered at the origin, the potential at a point $\mathbf{r} = (r > a, \theta, \phi)$ is

$$\Phi(\mathbf{r}) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \alpha_{\ell m} k_{\ell}(\kappa r) Y_{\ell m}(\theta, \phi) \quad (5)$$

with the multipole coefficients defined as a sum over the locations of the charges

$$\alpha_{\ell m} = 8\pi \sum_{\mathbf{r}_i=(r'_i, \theta'_i, \phi'_i)}^N q_i i_{\ell}(\kappa r'_i) Y_{\ell, -m}(\theta'_i, \phi'_i). \quad (6)$$

B. Poisson Boltzmann

A well-established electrostatics model which includes the solvent implicitly is the Poisson-Boltzmann equation.²⁵ The equation describes the electrostatic potential of a solvated system at the equilibrium state. Generalizing Gauss' law in a constant dielectric to a medium whose dielectric $\epsilon(\mathbf{r})$ is a spatially varying function, we get

$$-\nabla \cdot \epsilon(\mathbf{r}) \nabla \Phi(\mathbf{r}) = \rho(\mathbf{r}). \quad (7)$$

The functional form of $\epsilon(\mathbf{r})$ can be approximated by appealing to statistical mechanics. At a coarse enough level, it is reasonable to assume that the ions in the solution will be Boltzmann distributed. To obtain the Poisson-Boltzmann equation, we assume that the energy of the ion is equal to the charge times potential,⁹ giving

$$-\nabla \cdot \epsilon(\mathbf{r}) \nabla \Phi(\mathbf{r}) + \kappa^2 \sinh \Phi(\mathbf{r}) = \rho(\mathbf{r}). \quad (8)$$

The inverse Debye length κ , leads to an exponential decay in the field strength as a function of radial distance. For 1:1 electrolytes, the screening in SI units is $\kappa^2 = Ie^2/\epsilon k_B T$ where I is the ionic strength. This is a nonlinear, second-order elliptical partial differential equation. It can be linearized by assuming $e\Phi$ is small with respect to $k_B T$. By taking only the linear term in the expansion of the hyperbolic sine ($\sinh \Phi \approx \Phi$) the differential form of the Yukawa potential is recovered. That the linearized form of the Poisson-Boltzmann equation gives the Yukawa potential suggests that the Yukawa potential will serve as a good basis set. In this paper we solve the Poisson-Boltzmann equation with the Adaptive Poisson-Boltzmann Solver (APBS).²⁶ The solver uses an adaptive finite element mesh to model both the protein and an ionic solvent. The solver requires that the medium is partitioned into piecewise continuous regions of constant dielectric; typically these regions are the protein volume and the solvent.

III. METHODS

A. Data preparation

We examine three different proteins, lysozyme (1E8L),²⁷ ovalbumin (1OVA),²⁸ and human serum albumin (1E7H).²⁹ These three proteins were chosen as ideal models for their general conformational stability, globular geometry, and lack of a strong binding sites. The extension of the model to handle more complex interactions is discussed later. For each protein, we translate the atomic coordinates so the center of mass is at the origin. To simulate protonation states for a given pH, we first compute a pKa value for each charged residue using PROPKA.^{30,31} PDB2PQR,^{32,33} the conventionally used software for allocating partial charges at a given pH, only assigns states that are fully protonated or unprotonated. To titrate a continuous partial charge over the pH range we used the pKa value on the residue

$$v_i(\text{pH}) = \frac{10^{\text{pKa}_i - \text{pH}}}{1 + 10^{\text{pKa}_i - \text{pH}}}, \quad (9)$$

and assigned³⁴ the charge on the residue as

$$q_i^{\text{acid}} = v_i - 1, \quad (10)$$

$$q_i^{\text{base}} = v_i. \quad (11)$$

Using APBS we determined the full, nonlinear electrostatic potential Ψ from the Poisson-Boltzmann equation. For the dielectric constants, we used $\epsilon_p = 2$ inside the protein and $\epsilon_s = 78$ for the solvent. The ionic strength was $I = 150$ mM corresponding to Debye length of 8 Å.

We define a protein's diameter σ by the smallest bounding sphere that circumscribes the protein. All points inside a spherical shell of radius $a = 1.05\sigma$ are discarded. In this work we are only interested in reproducing the mid and far electrostatic potentials, so the series expansion presented in Sec. III is sufficient. Indeed, convergence is not even guaranteed at distances less than σ , nor is it completely accurate at the boundary when truncated to a finite order.

We used a fixed grid length with 161 points in each spatial direction. Using four levels of depth in the grid hierarchy

(APBS $n_{lev} = 4$) the number of grid dimensions must be a multiple of $n = c 2^{4+1} + 1$. Using fewer points resulted in numerical difficulty in calculating the multipoles, which we discuss in Sec. III B. The dielectric boundary between the protein and the solvent was calculated by taking the union of solvent-sized spheres which do not overlap with biomolecular atoms. This “free-volume” surface was then smoothed by nine-point harmonic averaging to reduce sensitivity to the grid parameters (`srfm=smol`).³⁵ The potential was calculated over two grids to improve accuracy. The coarse grid length was set to 20σ and the finer grid length spanned a distance of 6σ , with the Dirichlet boundary conditions of multiple Debye-Hückel spheres (`bconf=mdf`). Prior to computation, hydrogen atoms were added to the full atomic coordinates with the PARSE parameter set. The field was computed at 120 equally spaced intervals over the pH range of 3–12.

B. Calculation of APBS multipoles

Since a typical Poisson-Boltzmann calculation yields about 4×10^6 grid points for the electrostatic potential, it is inefficient to fit the potential of a set of effective charges to this grid. In contrast, only a few dozen multipole moments are needed to reproduce the same field. Our *modus operandi* then is to fit the multipole moments of the effective charges to the multipole moments of the original field. This improves the fitting time by several orders of magnitude as more function evaluations can be made.

For convenience in notation, we convert the double index of multipole coefficients and spherical harmonics to a single index viz. $\alpha_k = \alpha_{\ell(\ell+1)+m}$, e.g., $\alpha_0 \equiv \alpha_{0,0}$, $\alpha_1 \equiv \alpha_{1,-1}$, $\alpha_2 \equiv \alpha_{1,0}$, $\alpha_3 \equiv \alpha_{1,1}$, $\alpha_4 \equiv \alpha_{2,-1}$. Boldface notation will represent an ordered vector of these coefficients $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_L]$ or the charge magnitudes $\mathbf{q} = [q_1, q_2, \dots, q_N]$. Since we truncate the expansion we denote the total number of terms as $L = (\ell_{MAX} + 1)^2$. ℓ_{MAX} is the value of ℓ where the series is truncated. Unless otherwise noted, we took $\ell_{MAX} = 8$ corresponding to $L = 81$. In practice, however, not all of the multipole terms are unique since $\alpha_{\ell,-m} = (-1)^m \alpha_{\ell,m}^*$.

We start by computing the multipole coefficients α of Eq. (6), from the original APBS field Ψ . If the field was a known collection of discrete point charges, the multipole coefficients would be calculated directly from Eq. (5), the dis-

FIG. 1. Multipole fit quality for ovalbumin at an ionic strength of $I = 150$ mM as a function of pH. Fit quality χ is defined by the difference of the target APBS field Ψ to the multipole field Φ via Eq. (14). Shown in descending order are the fits for $\ell_{MAX} = 0, 1, 2, \dots, 8$.

crete version of the Yukawa potential. Alternatively, since our basis functions form an orthonormal set $\int_V Y_{\ell,m}^* Y_{\ell',m'} dV = \delta_{\ell,\ell'} \delta_{m,m'}$, given the potential at all points we could integrate to find the coefficients. In general however, we are only given the value of Ψ at a discrete set of m points. We can determine α , the best fit coefficients to the field, as a solution to the over-determined set of equations

$$\mathbf{M}\alpha = \Psi, \quad (12)$$

where $\mathbf{M}_{ij}$ is the contribution of the i th multipole moment to the j th point in space (r_j, θ_j, ϕ_j)

$$\mathbf{M}_{ij} = k_i(\lambda r_j) Y_i(\theta_j, \phi_j). \quad (13)$$

Given a target electrostatic potential Ψ and an attempted reproduction Φ , we define the quality of fit between the potentials as

$$\chi(\Psi, \Phi) = \frac{\int \|\Psi(\mathbf{r}) - \Phi(\mathbf{r})\|^2 d^3\mathbf{r}}{\int \|\Psi(\mathbf{r})\|^2 d^3\mathbf{r}}. \quad (14)$$

Values of $\chi = 0$ indicate a perfect fit. Using this measure we found that the potential from the multipole moments did an excellent job of reproducing the original field for sufficiently large ℓ_{MAX} , see Figures 1 and 2. It is evident that, even including terms up to the quadrupole moment (red line, $\ell = 2$), the expansion fails to capture the field for all pH values. This has important ramifications for modeling the electrostatic potential by a low-order multipole expansion. For

FIG. 2. Multipole fit quality for human serum albumin (left) and lysozyme (right) at $I = 150$ mM. Figure labels are the same as Fig. 1.

many pH ranges, a significant portion of the potential has a stronger orientational dependence than can be captured by the lower multipoles alone. This is especially true near the isoelectric point for each protein. At this pH the protein is electroneutral. Here the degree of anisotropy is high since the monopole term ($\ell = 0$) vanishes. However, a sufficiently large enough value of ℓ_{MAX} can be chosen to reduce the error by an arbitrary amount. Unsurprisingly, the same fits using only the spherical harmonics without the Bessel functions, Eq. (2), were insufficient to capture the ionic decay.

When integrated over the proper domain, the importance of each multipole in reproducing the original field is proportional to the magnitude

$$\mathbf{w}_\alpha = [|\alpha_1|, |\alpha_2|, \dots, |\alpha_L|]. \quad (15)$$

However, when integrated over a partial region of space the contribution from each basis function is not necessarily orthogonal. In our calculation of the multipole coefficients we found that the contribution was sufficiently orthogonal. See Appendix B for more detailed discussion of this issue.

C. Determining the effective charge magnitude and location

In both the effective charge and multipole representation, (Eqs. (3) and (6), respectively) the potential is linear in charge magnitude. We can reduce the number of free variables per effective charge from four to three by setting the charge magnitudes to the linear fit

$$\mathbf{F}\mathbf{q} = \boldsymbol{\alpha}, \quad (16)$$

where $\mathbf{F}_{ij}$ is the contribution of the i th charge to the j th multipole moment (Eq. (6))

$$\mathbf{F}_{ij} = i_j(kr'_i)Y_j(\theta'_i, \phi'_i). \quad (17)$$

We experimented with an additional constraint, the magnitude of the total charge

$$q_M - \sum_{i=1}^N q_i = 0. \quad (18)$$

We added this constraint to the set of charge magnitude equations with a Lagrange multiplier. The total charge q_M was constrained by the monopole term, the magnitude of a single fixed charge at the origin. We found that constraining the net charge did not have a large impact on the quality of the resulting fits. Unlike the unscreened case, only large anisotropic portions of the field are significant at sizable distances. Therefore, two different effective charge distributions may reproduce the same field even if the net charges are not the same for each.

Since the relationship between the effective charge locations and $\boldsymbol{\alpha}$ is highly nonlinear, the converged solution depends on the initial locations chosen. Therefore, a prudent choice of initial conditions will lead to faster convergence rates. The k atomic charge locations from the original protein coordinates suggest such a starting point. The number of fitted charges is less than than atomic charge count, thus a

subset of these charges must be chosen. We start by computing the contribution to the multipole moments of the k atomic charges via Eq. (17). By a singular value decomposition, we retain the largest eigenvector and place charges at the N (absolute) largest components of that vector. In this respect, we start with the mode that contributes the most to the potential while leaving the other modes to be fit during the nonlinear portion of the algorithm. Once the charged residues have been selected, the initial coordinate for the effective charge is placed on the α -carbon atom.

At this point there are no more linear variables to optimize over as the relationship between the effective charge locations and multipole moments is nonlinear. Before fitting the positions we take note of several factors. There is a danger in letting the effective charges take arbitrary positions without additional restrictions. The most obvious issue is the placement of one of the charges outside the spherical region circumscribing the protein. In this case the expansion is no longer valid, hence we define a bias function to be added to the fitting weight to prevent this

$$b_1(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = \sum_i^N (r_i - \sigma)^2 \Theta(r_i - \sigma), \quad (19)$$

where Θ is the Heaviside step function.

Another more subtle issue is the creation of an artificial dipole. As two effective charges get close with roughly opposite charges, the effect on the potential is scaled not by their magnitudes, but on the factor of qd , their charge times their separation. As long as qd remains constant, the dipole charge can fluctuate wildly yet still contribute to the potential equally. When left unchecked, this can create charge magnitudes that vary over several orders of magnitude in the same protein. Note that even setting the net charge to a constant (or fit with a Lagrange multiplier) will not remedy the situation since the net charge on the artificial dipole is approximately zero. We therefore define a constraint in our fitting function that penalizes the formation of dipoles smaller than a characteristic length. With a minimum distance of d_{min} we set a second bias function as

$$b_2(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = \sum_i^N \sum_{j \neq i}^N (r_i - r_j)^2 \Theta(|r_i - r_j| - d_{\text{min}}). \quad (20)$$

At large values d_{min} we noticed a degradation in the fit quality. We choose a value of $d_{\text{min}} = 0.2 \text{ \AA}$ as smaller distances offered no additional improvement but increased the convergence time of the fits.

With the multipole moments of the effective charge defined as $\boldsymbol{\beta}$, we seek to minimize

$$f(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) = \mathbf{w}_\alpha \cdot |\boldsymbol{\alpha} - \boldsymbol{\beta}|^T + b_1 + b_2. \quad (21)$$

The function was minimized with the help of a downhill simplex algorithm. A maximum 10^4 iterations were performed to achieve convergence, which we defined as having a tolerance of $|f_{i+1} - f_i| \leq 10^{-5}$. While undoubtedly fast, the simplex method can potentially become trapped in local minima. To improve the fits, we ran an annealing procedure over the pH range. The algorithm was repeated each time with initial conditions chosen randomly from one of the best fit solutions around a window of $\Delta\text{pH} \pm 1.0$. Since each pH value could

be run independently, an iteration of the annealing procedure was trivially parallelizable. The parallel runs utilized Biowulf, a heterogeneous high-performance cluster at the National Institutes of Health. Each individual calculation required 2–6 GB of memory and several hours for each call to APBS. Additionally, several hours were needed for the initial non-linear fit (with negligible memory cost). Convergence of the annealing procedure was much quicker since it started from a local minima of a perturbed system, typically it completed in less than 10 min. While the entire procedure requires an initial upfront computational cost (which is greatly reduced when parallelized over a pH range), once the effective charges have been solved for they provide substantial gains in the computation of the interaction potential.

IV. RESULTS

For each protein we fit $N = 12$ charges with $\ell_{\text{MAX}} = 8$. The field produced by the effective charges was found by the discrete sum via Eq. (3) and the quality of fit to the target potential field Ψ was calculated from Eq. (14). Even a perfect fit of our minimization function ($f = 0$) would not imply that the fields match ($\chi = 0$), only that the multipole moments of the two fields are close to a given order. However, we should expect that a good match in the multipole moments should correspond to a high quality of fit of the respective fields. As an example, consider Figure 3 which illustrates the difference of the multipole coefficients of the original field versus the effective charges. The top panel explores the weighted contribution of each multipole coefficient to the potential. The isoelectric point is clearly visible around $\text{pH} = 4.4$ where the monopole moment (red) vanishes and the other coefficients become more important. The corresponding multipole coefficients generated by the effective charges do an excellent job at matching these proportions as shown in the bottom panel of the same figure. We see that the fitting error χ shown in Figure 4 visually correlates the error of the multipole fits to that of the effective charge model. Assuming convergence, we found a linear relation between the quality of fit of the

FIG. 3. Stacked area plots of the multipole coefficients α of the original field (top) and the multipole coefficients of the effective charges β (bottom) of ovalbumin. For comparison purposes across different pH values, the vectors have been normalized to unity. Starting from the bottom, each color is a different value of ℓ with the lighter shades representing the values of m (red $\rightarrow \ell = 0$, green $\rightarrow \ell = 1$, blue $\rightarrow \ell = 2$, ...). While there is some uncertainty near the isoelectric point ($\text{pH} = 4.4$), the effective charges do a good job at capturing the multipole moments.

FIG. 4. (Top) Quality of fit χ (Eq. (14)) of the original field to the field generated by 12 effective charges for ovalbumin at $I = 150$ mM as a function of pH. The error is shown on a log scale, with the red dashed line indicating a 10% error. The solid line indicates the quality of fit for a corresponding multipole generated field which essentially acts as a lower bound to the error. (Bottom, left and right) Quality of fit χ for human serum albumin and lysozyme, respectively. Axes labels and scales are the same for all plots.

multipoles versus the effective charges. The fitting error for the other two proteins, illustrated in the bottom panels of Figure 4, also show a low error for almost all pH values. Naturally the quality of the effective charge fits followed the same trend as the multipoles; at best the effective charges can only match the input potential generated by the multipoles.

Once the fitting function f has been minimized (Eq. (21)), the multipole moments for the effective charges should reproduce the original field. The fits obtained are reasonable, that is small χ , but the effective charge coordinates do not always vary continuously over small changes in the pH. Due to the nonlinear relation between the effective charge position and the basis functions, small conformational changes in the partial charges location or magnitude led to large changes in the positions of the effective charges. In a paper by Gabdoulline and Wade,¹⁶ it was noted that an effective charge model displayed sensitivity to different conformations of the same protein. An eigen-analysis of the effective charges over multiple conformations showed that there was a set of generalized charges, linear combinations of the effective charges, that was more robust than any individual conformation. While not investigated in this paper, a charge distribution that reproduces the electrostatic potential and is robust to conformational changes could be an improvement to the method we have presented.

An ionic strength of $I_{\text{ref}} = 150$ mM was chosen as a physiological relevant strength for all calculations. To illustrate the applicability of the model to lower ionic strengths, the calculations were repeated at an ionic strength of $I_{\text{low}} = 15$ mM. In this regime we expect better performance from the model since the linearization of the PB equation into a Yukawa potential becomes less dependent on the screening. Consequently, the initial stage of multipole fits via Eq. (12) are more accurate. Since this multipole representation captures more detail of the original potential the corresponding fits of the effective charges should be better. Indeed, at these lower ionic strengths we found that the error in the final fit was always better and approximately one third of the reference ionic strength. The precise dependence of the fitting error against ionic strength is likely nonlinear, especially

FIG. 5. Logarithmic dependence of the average error at fixed $\ell_{\text{MAX}} = 8$ for a variable number of effective charges N . The error is the same as Figure 4, only averaged over all pH values.

at ionic strengths $> I_{\text{ref}}$, but was not investigated further in this work.

The relation between the number of effective charges and the averaged quality of fit is shown in Figure 5. Generally speaking, each effective charge adds logarithmic returns on the quality of fit at the same rate for each protein. Heuristically, this provides a nice metric for determining the number effective charges necessary for a given level of accuracy. We found that $N = 12$ effective charges gave a good balance of fit accuracy to charge number since as the average error for all studied proteins was significantly less than 10%.

A schematic illustration of the effective charges against the protein structure at the isoelectric point is shown in Figures 6–8. Each picture represents both the partial charges on the residues used as input to APBS and the fitted effective charges as comparison. It is clear that the number of charges needed to reproduce the input potential has been reduced. This reduction is even more impressive when interaction potentials are calculated as the energy simply becomes an additive pairwise sum.

A. Interaction energies and virial coefficients

We first compare the prediction made by APBS and our effective charges to the electrostatic interaction energy. The electrostatic potential of mean force $\Delta\Delta U_\phi$ for two proteins in an inhomogeneous medium is decomposed into the energy of cavity creation for each protein $\Delta U_\phi^{\text{p1}}$ and the interaction of the charged complex

$$\Delta\Delta U_\phi = \Delta U_\phi^{\text{complex}} - \Delta U_\phi^{\text{p1}} - \Delta U_\phi^{\text{p2}}. \quad (22)$$

For a single configuration at a fixed pH, each calculation of the interaction energy took several hours as it involved six separate calls to APBS. In contrast, a near instantaneous result was obtained with the effective charges since the calculation was a simple summation of pairwise contributions in a homogeneous medium. There is excellent agreement with the two methods for all proteins considered at a distance of $L = 1.25\sigma$, illustrated in Figure 9. At or near contact $L \leq \sigma + \kappa/2$, the fits tend to be less accurate as the granularity of the $N = 12$ effective charge approximation is insufficient to capture the finer details of potential governed by the

FIG. 6. (Top) Locations of the charged residues for human serum albumin at its isoelectric point $\text{pH} \approx 4.6$. Positive/negative charges are shown in blue/red, and the volume of the illustrated sphere is proportional to the charge magnitude. Secondary structure is shown in the gray cartoon model. (Bottom) Best fit effective charge model $N = 12$, $\ell_{\text{MAX}} = 8$ at the same pH. Not all 12 charges can be shown simultaneously for any fixed orientation.

PB equation. At this range, however, other forces may play a more dominant role and an alternate representation may be appended to the Yukawa potential. Our model is applicable over ranges where the resolution of the protein can reasonably be described by these electrostatic potentials alone.

In general, since the interaction energies at various orientations are well described by the effective charges we can compute a more complex statistic. The second virial coefficient B_{22} , describes the pair-interaction between two molecules. Normally the computation follows by integrating over an average of the potential of mean force. While impossible to calculate explicitly with APBS, it is straight-forward with a set of effective charges

$$B_{22} = -2\pi \int \langle (e^{-\Phi_{12}/k_B T} - 1) r^2 dr \rangle, \quad (23)$$

where Φ_{12} is the pair potential of two molecules and the brackets indicate an average over all orientations at a fixed distance.³⁶ In our case, due to the anisotropy of the interaction potential, we express the full angular dependency by integrating over all angular coordinates.³⁷ We center one protein at the origin and translate the other to the polar coordinates (r, θ, ϕ) . The displaced protein is then rotated at Euler angles α, β, γ . The second virial coefficient is³⁸

$$B_{22} = \frac{-1}{16\pi^2} \int_0^{2\pi} \int_0^\pi \int_0^{2\pi} \int_0^{2\pi} \int_0^\pi \int_0^\infty (e^{-\Phi_{12}/k_B T} - 1) \times r^2 \sin \theta \sin \beta dr d\theta d\phi d\alpha d\beta d\gamma. \quad (24)$$

FIG. 7. (Top) Locations of the charged residues for lysozyme at its isoelectric point $\text{pH} \approx 11.4$. (Bottom) Effective charge model. Representation is the same as Figure 6.

At a distance of σ , the maximum possible value of steric overlap, we split the integral into two parts. When the proteins overlap the interaction energy is infinite. This contribution to B_{22} is equivalent to that of a hard-sphere and can be integrated out. While this does not take into account any geometrical anisotropy at length scales less than σ , we cannot go any smaller as our expansion is not valid at these distances. By grouping all of the remaining angular integration variables as Ω

$$B_{22} = \frac{2\pi\sigma^3}{3} + \frac{-1}{16\pi^2} \int_{\Omega} [(e^{-\Phi_{12}/k_B T} - 1)r^2 dr] d\Omega$$

$$= B_{HS} + B_E. \quad (25)$$

The first term B_{HS} represents the positive excluded volume contribution and the second term B_E represents the positive or negative electrostatic contributions. Using a hybrid combination of adaptive quadrature integration and Monte Carlo sampling, we computed B_E over a log-spacing from $r = \sigma$ to 5σ . A cutoff of 5σ was found to be sufficient as the exponential damping decayed the potential to a negligible value. We plot the results in excess of the hard sphere con-

FIG. 8. (Top) Locations of the charged residues for ovalbumin at its isoelectric point $\text{pH} \approx 4.3$. (Bottom) Effective charge model. Representation is the same as Figure 6.

tribution B_{22}/B_{HS} , shown in Figure 10. Near the isoelectric point the electrostatic contribution is small as it is effectively an orientational average. However, any particular configuration of the proteins, as shown in the interaction energies of Figure 9, may still contribute to the potential. This difference may be important for numerical studies of protein-protein interactions where the entire conformational space is sampled.

We note that this computation of the second virial coefficient only takes into account two very specific classes of protein-protein interaction: volume exclusion and electrostatic interactions. Hydrophobic effects, conformational flexibility, polarizability, and other short range solvent-mediated interactions will play a large role in determining the true value of B_{22} relative to the measured empirical values.¹⁷ Nevertheless, it is useful to isolate the electrostatic contribution to make comparisons against other interactions and estimate the role it

FIG. 9. Interaction energy $\Delta\Delta U_\phi$ of two identical molecules separated by a distance of 1.25σ along the x-axis. The effective charges were computed from a single molecule. The interaction energy calculated from APBS, shown as dashed black line, was computed from Eq. (22) took several hours per configuration. In contrast, the computational cost of the effective charge model was negligible. Even at the isoelectric point where monopole moment vanishes, the effective charges do a good job of reproducing the interaction potential.

plays in the overall potential. The qualitative trends of pH dependence in experimental measurements of the second virial coefficient in lysozyme^{39–43} and ovalbumin⁴⁴ are similar, suggesting the significance of the electrostatic contribution.

V. GENERAL DISCUSSION

We have shown in this paper that a subset of effective charges, smaller than the original number of charged groups, can reproduce the anisotropic electrostatic potential of a protein in an ionic solution. This was done by allowing the charges to take arbitrary positions within the protein and fit instead the multipole moments of each distribution instead of the original field. The only restriction we imposed on the effective charges was a bias term to combat the degeneracy a dipole creates in the fitting process.

The method was robust enough to handle the proteins presented, which were roughly spherical and conformationally stable. To extend the fitting procedure to handle a substantially anisotropic geometry would require an expansion of the potential inside the bounding sphere. This presents a new line of research to improve the quality of the fits, where

FIG. 10. Scaled electrostatic contribution to the second virial coefficient B_{22}/B_{HS} for the three studied molecules. Values at one signify no excess electrostatic contribution to the integral. The calculations were performed at ionic strength $I = 150$ mM using $N = 12$, $\ell_{MAX} = 8$. The isoelectric points are marked with a solid dash for each molecule.

the potential would be fit within the spherical bounding shell but still outside the molecular surface. The calculation of the excess electrostatic virial coefficient would be more accurate in this case since the excluded volume would depend on this geometry as well. Proteins that form a strong complex may also present unique difficulties with this model without supplementation. These interactions are much larger in magnitude than the non-specific potential, with typically concave geometries and forces that cannot be fully accounted for in the Poisson-Boltzmann equations.

In many cases, the potential can be separated into additive components with the electrostatic portion well approximated by the Poisson-Boltzmann equation. The effective charge model is advantageous as it offers a fast evaluation of the potential at mid to far distances even at the isoelectric point. Effective charges, used in conjunction with other models for the close range potential, can serve as a good coarse-grained approximation for large scale simulations.

ACKNOWLEDGMENTS

We would like to thank Allen Minton for his help and support in the production and editing of this paper. The work was supported in part by the intramural program of NIDDK at the NIH. This study utilized the high-performance computational capabilities of the Helix/BioWulf Systems at the National Institutes of Health, Bethesda, MD (<http://helix.nih.gov>).

APPENDIX A: MODIFIED SPHERICAL HARMONICS

As a reference we provide complete formulas for the spherical harmonics and the modified spherical Bessel functions. Standard nomenclature denotes the modified spherical Bessel function $i_n(x)$. To avoid confusion we write the imaginary unit $i \equiv \sqrt{-1}$ out explicitly in the formula below. Let $P_n(x)$ denote the Legendre polynomial of degree n and the associated Legendre functions $P_{n,m}(x)$ be defined by Rodrigues' formula

$$P_{n,m}(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_n(x). \quad (A1)$$

The spherical harmonics are defined as

$$Y_{\ell,m}(\theta, \phi) = P_{n,|m|}(\cos\theta) e^{\sqrt{-1}m\phi} \sqrt{\frac{2n+1}{4\pi}} \sqrt{\frac{(n-|m|)!}{(n+|m|)!}}. \quad (A2)$$

The modified spherical Bessel functions are defined in terms of the Bessel function of the first kind J_ν

$$J_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m (x/2)^{2m+\nu}}{\Gamma(\nu+m+1)m!}. \quad (A3)$$

The functions i_n and k_n used in Eqs. (4)–(6) is given by

$$i_n(r) = \sqrt{\frac{\pi}{2r}} I_{n+1/2}(r), \quad (A4)$$

$$k_n(r) = \sqrt{\frac{\pi}{2r}} K_{n+1/2}(r), \quad (\text{A5})$$

with the functions $I_\nu(r)$ and $K_\nu(r)$ defined

$$I_\nu(r) = (\sqrt{-1})^{-\nu} J_\nu(\sqrt{-1}r), \quad (\text{A6})$$

$$K_\nu(r) = \frac{\pi}{2 \sin \nu\pi} (I_{-\nu}(r) - I_\nu(r)). \quad (\text{A7})$$

APPENDIX B: ALTERNATE FITTING WEIGHTS

Integrated over the proper domain any two orthogonal basis functions with different indices will be zero. Integration over partial regions of the domain, under-sampling or coarse grid representations of the potential may lead to non-orthogonality. In this case there is a new orthogonal basis set where the contribution of the original basis functions form some linear combination. To check to see if this was an issue for our calculations, we computed a generalized multipole. This was estimated via a singular value decomposition (SVD),

$$\mathbf{U}\mathbf{\Lambda}\mathbf{V} = \mathbf{A}\mathbf{F}, \quad (\text{B1})$$

where $\mathbf{A}_i = \boldsymbol{\alpha}$. In the decomposition, $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_L]$. With m points in the original field, the matrices have size $\mathbf{U}_{[m,L]}$ and $\mathbf{V}_{[L,m]}$. Computing the SVD for a large number of points is computationally taxing, and ultimately unnecessary since only a few of the largest eigenvectors are dominant. We only compute a truncated portion retaining L eigenvectors using fast stochastic SVD methods,⁴⁵ $\mathbf{U}_{[m,L]} \rightarrow \mathbf{U}'_{[L,L]}$. An alternate fitting weight computed from the multipoles is

$$\mathbf{w}_\lambda = \mathbf{U}'^T \boldsymbol{\lambda}. \quad (\text{B2})$$

Ideally, our fitting function should be weighted to this linear combination.

Over very coarse grids, where the grid spacing was comparable to the protein diameters, we noticed a significant mixing in the generalized multipoles; the vectors in $\mathbf{U}$ were not orthogonal. For smaller, more reasonable grid lengths, $\mathbf{w}_\alpha$ and $\mathbf{w}_\lambda$ gave the same predictions for each individual multipole. However, compared to $\mathbf{w}_\alpha$, we found that $\mathbf{w}_\lambda$ underestimated the weight of the monopole term and gave disproportionate weight to the high-orders. In addition, this weighting suffered from numerical noise due to the stochastic nature of the fast-SVD algorithm. Ultimately, for our grid lengths the quality of the fits were approximately the same in both cases so the simpler weighting given by $\mathbf{w}_\alpha$ was used.

¹A. P. Minton, "The effective hard particle model provides a simple, robust, and broadly applicable description of nonideal behavior in concentrated solutions of bovine serum albumin and other nonassociating proteins," *J. Pharm. Sci.* **96**(12), 3466–3469 (2007).

²G. Franzese, G. Malescio, A. Skibinsky, S. V. Buldyrev, and H. E. Stanley, "Generic mechanism for generating a liquid-liquid phase transition," *Nature (London)* **409**(6821), 692–695 (2001).

³B. Lu and A. R. Denton, "Phase separation of charge-stabilized colloids: A Gibbs ensemble Monte Carlo simulation study," *Phys. Rev. E* **75**(6), 061403 (2007).

⁴A. Panagiotopoulos, "Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble," *Mol. Phys.* **61**(4), 813–826 (1987).

⁵T. Hoppe and J.-M. Yuan, "Protein folding with implicit crowders: A study of conformational states using the Wang-Landau method," *J. Phys. Chem. B* **115**(9), 2006–2013 (2011).

⁶P. Camp, J. Shelley, and G. Patey, "Isotropic fluid phases of dipolar hard spheres," *Phys. Rev. Lett.* **84**(1), 115–118 (2000).

⁷G. Ganzenmüller, G. N. Patey, and P. J. Camp, "Vapour-liquid phase transition of dipolar particles," *Mol. Phys.* **107**(4–6), 403–413 (2009).

⁸R. Bharadwaj, A. Windemuth, S. Sridharan, B. Honig, and A. Nicholls, "The fast multipole boundary element method for molecular electrostatics: An optimal approach for large systems," *J. Comput. Chem.* **16**(7), 898–913 (1995).

⁹B. Z. Lu, Y. C. Zhou, M. J. Holst, and J. A. McCammon, "Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications," *Comm. Comp. Phys.* **3**(5), 973–1009 (2008).

¹⁰M. J. Schnieders, N. A. Baker, P. Ren, and J. W. Ponder, "Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum," *J. Chem. Phys.* **126**(12), 124114 (2007).

¹¹R. P. Leavitt, "An irreducible tensor method of deriving the long-range anisotropic interactions between molecules of arbitrary symmetry," *J. Chem. Phys.* **72**(6), 3472 (1980).

¹²L. F. Greengard and J. Huang, "A new version of the fast multipole method for screened coulomb interactions in three dimensions," *J. Comput. Phys.* **180**(2), 642–658 (2002).

¹³J. Makino, "Yet another fast multipole method without multipoles—Pseudo-particle multipole method," *J. Comput. Phys.* **151**(2), 910–920 (1999).

¹⁴C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model," *J. Phys. Chem.* **97**(40), 10269–10280 (1993).

¹⁵T. R. Stouch and D. E. Williams, "Conformational dependence of electrostatic potential-derived charges: Studies of the fitting procedure," *J. Comput. Chem.* **14**(7), 858–866 (1993).

¹⁶R. R. Gabdoulina and R. C. Wade, "Effective charges for macromolecules in solvent," *J. Phys. Chem.* **100**(9), 3868–3878 (1996).

¹⁷A. H. Elcock and J. A. McCammon, "Calculation of weak protein-protein interactions: The pH dependence of the second virial coefficient," *Biophys. J.* **80**(2), 613–625 (2001).

¹⁸D. A. Beard and T. Schlick, "Modeling salt-mediated electrostatics of macromolecules: the discrete surface charge optimization algorithm and its application to the nucleosome," *Biopolymers* **58**(1), 106–115 (2001).

¹⁹Q. Zhang, D. A. Beard, and T. Schlick, "Constructing irregular surfaces to enclose macromolecular complexes for mesoscale modeling using the discrete surface charge optimization (DISCO) algorithm," *J. Comput. Chem.* **24**(16), 2063–2074 (2003).

²⁰S. Alexander, P. M. Chaikin, P. Grant, G. J. Morales, P. Pincus, and D. Hone, "Charge renormalization, osmotic pressure, and bulk modulus of colloidal crystals: Theory," *J. Chem. Phys.* **80**(11), 5776 (1984).

²¹C. Alvarez and G. Téllez, "Screening of charged spheroidal colloidal particles," *J. Chem. Phys.* **133**(14), 144908 (2010).

²²A. P. dos Santos, A. Diehl, and Y. Levin, "Colloidal charge renormalization in suspensions containing multivalent electrolyte," *J. Chem. Phys.* **132**(10), 104105 (2010).

²³J. D. Jackson, *Classical Electrodynamics*, 3rd ed. (Wiley, 1999).

²⁴M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1972).

²⁵M. E. Davis and J. A. McCammon, "Electrostatics in biomolecular structure and dynamics," *Chem. Rev.* **90**(3), 509–521 (1990).

²⁶N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proc. Natl. Acad. Sci. U.S.A.* **98**(18), 10037–10041 (2001).

²⁷H. Schwalbe, S. B. Grimshaw, A. Spencer, M. Buck, J. Boyd, C. M. Dobson, C. Redfield, and L. J. Smith, "A refined solution structure of hen lysozyme determined using residual dipolar coupling data," *Protein Sci.* **10**(4), 677–688 (2001).

²⁸P. E. Stein, A. G. Leslie, J. T. Finch, and R. W. Carrell, "Crystal structure of unclaved ovalbumin at 1.95 Å resolution," *J. Mol. Biol.* **221**(3), 941–959 (1991).

- ²⁹A. A. Bhattacharya, T. Grüne, and S. Curry, "Crystallographic analysis reveals common modes of binding of medium and long-chain fatty acids to human serum albumin," *J. Mol. Biol.* **303**(5), 721–732 (2000).
- ³⁰H. Li, A. D. Robertson, J. H. Jensen, N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Very fast empirical prediction and rationalization of protein pKa values," *Proteins* **61**(4), 704–721 (2005).
- ³¹D. C. Bas, D. M. Rogers, and J. H. Jensen, "Very fast prediction and rationalization of pKa values for protein-ligand complexes," *Proteins: Struct., Funct., Bioinf.* **73**(3), 765–783 (2008).
- ³²T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker, "PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations," *Nucleic Acids Res.* **32**(Web Server issue), W665–W667 (2004).
- ³³T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker, "PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations," *Nucleic Acids Res.* **35**(Web Server issue), W522–W525 (2007).
- ³⁴C. Tanford, *Physical Chemistry of Macromolecules* (John Wiley & Sons Inc, 1961).
- ³⁵R. E. Brucoleri, J. Novotny, M. E. Davis, and K. A. Sharp, "Finite difference Poisson-Boltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing," *J. Comput. Chem.* **18**(2), 268–276 (1997).
- ³⁶B. H. Zimm, "Application of the methods of molecular distribution to solutions of large molecules," *J. Chem. Phys.* **14**(3), 164 (1946).
- ³⁷B. Neal, "Why is the osmotic second virial coefficient related to protein crystallization?," *J. Cryst. Growth* **196**(2–4), 377–387 (1999).
- ³⁸J. O. Hirschfelder, C. F. Curtiss, and R. B. Bird, *The Molecular Theory of Gases and Liquids* (Wiley-Interscience, 1964).
- ³⁹V. Le Brun, W. Friess, T. Schultz-Fademrecht, S. Muehlau, and P. Garidel, "Lysozyme-lysozyme self-interactions as assessed by the osmotic second virial coefficient: Impact for physical protein stabilization," *Biotechnol. J.* **4**(9), 1305–1319 (2009).
- ⁴⁰O. D. Velev, E. W. Kaler, and A. M. Lenhoff, "Protein interactions in solution characterized by light and neutron scattering: Comparison of lysozyme and chymotrypsinogen," *Biophys. J.* **75**(6), 2682–2697 (1998).
- ⁴¹P. M. Tessier, A. M. Lenhoff, and S. I. Sandler, "Rapid measurement of protein osmotic second virial coefficients by self-interaction chromatography," *Biophys. J.* **82**(3), 1620–1631 (2002).
- ⁴²D. F. Rosenbaum and C. F. Zukoski, "Protein interactions and crystallization," *J. Cryst. Growth* **169**(4), 752–758 (1996).
- ⁴³Y. U. Moon, C. O. Anderson, H. W. Blanch, and J. M. Prausnitz, "Osmotic pressures and second virial coefficients for aqueous saline solutions of lysozyme," *Fluid Phase Equilib.* **168**(2), 229–239 (2000).
- ⁴⁴A. C. Dumetz, A. M. Chockla, E. W. Kaler, and A. M. Lenhoff, "Effects of pH on protein-protein interactions and implications for protein phase behavior," *Biochim. Biophys. Acta* **1784**(4), 600–610 (2008).
- ⁴⁵N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," eprint [arXiv:0909.4061](https://arxiv.org/abs/0909.4061).